Learning Patterns for Information Extraction from Free Text

Conrad Plake, Jörg Hakenberg, Ulf Leser
Humboldt-Universität zu Berlin, 10099 Berlin, Germany
{plake,hakenberg,leser}@informatik.hu-berlin.de

Abstract

We describe a general approach to the task of information extraction from free text and propose methods for learning syntax patterns automatically from annotated corpora. We study the application of our approach to the extraction of protein-protein interactions from scientific texts. Based on this evaluation, we find that learning patterns outperforms techniques based on hand-crafted patterns.

1 Introduction

Information Extraction (IE) in the life science domain has become a big challenge for the textmining community. The amount of research articles in this field is immense and still exponentially growing. Today the citation index MEDLINE references more than 15 million abstracts. Such large amount of text rises the need for automatic text processing. This paper describes a general approach to the task of IE from free text. In this context we think of information as named entities and their relations. For a successful extraction of relations, many preliminary steps, such as *tokenization*, *part-of-speech* (POS) tagging and *named-entity recognition* (NER), have to be accomplished. We focus on the extraction of relations between entities in general and introduce methods for learning patterns from annotated training corpora. We evaluated our approach by extracting protein-protein interactions (PPI) from abstracts and full texts of research articles, a task widely studied during the last years [Friedman *et al.*, 2001; Blaschke and Valencia, 2002; Daraselia *et al.*, 2004]. We applied our methods to a set of 1000 sentences, taken from the BioCreAtIvE-Corpus [BioCreAtIvE, 2004], and to articles describing PPI stored in DIP [Xenarios *et al.*, 2001].

2 Methods

We followed two different approaches. The first was based on algorithms for pairwise sequence alignment to generate patterns and match them against new text. Huang *et al.* examined such methods to extract PPI from biomedical texts annotated with POS-information [Huang *et al.*, 2004]. Patterns were sequences of anchor words and used for alignment or as regular expressions, which can be further refined. By applying additional properties, *e.g.* limited wordgaps, a set of patterns was optimized using a genetic algorithm [Plake *et al.*, 2005].

In the second approach, we did not perform explicit pattern generation, but rather tried to find patterns in annotated texts. We modeled patterns as *finite state automatons* (FSA), whose structures were learned from scratch by genetic optimization. We distinguished between a deterministic and a probabilistic model. The first was a Mealy-FSA, that can easily be used for regular grammar text parsing. The second model was a Hidden Markov Model (HMM). For an observation sequence, *i.e.* a sentence, the most probable sequence of states was calculated with the Viterbi-algorithm [Viterbi, 1967]. Task-specific entities and word classes were each represented by a state – proteins, interactors and different POS tags. All transition- and emission probabilities were derived from fully annotated training sentences. Figure 1 shows representations of a learned pattern for each approach.

A training corpus of 1000 sentences has been manually inspected for any kind of PPI and 253 interactions were stored in a gold file to support automatic evaluation. We performed a 10-fold crossvalidation and compared the extraction accuracy of automatically learned patterns to a set of hand-crafted patterns. In every fold a random selection of two-thirds of the corpus was used for training and one-third for testing. For learning multiple patterns, we followed a separate-and-conquer strategy.

A second validation was performed on articles describing protein interactions stored in the DIP. The DIP contains nodes for known PPI, each including references to a research article describing the respective experimental findings. We randomly selected 297 PPI for the yeast organism and sought to find these interactions in the referenced literature. Again, we compared results of automatically learned patterns to hand-crafted ones.

3 Results

In this paper, we present preliminary results of our work. First we examined patterns, build by pairwise local alignment of example sequences. As a scoring scheme for substitution, we chose a modified Levenshtein measure. Aligning patterns to sentences resulted in 90% precision, or 60% recall respectively. Balance between precision and recall was adjusted by a threshold of the pattern generation algorithm proposed by Huang *et al.*, 2004. Patterns occurring less often than a specified threshold were ignored. When using the sequences as regular expressions, further optimized with wordgap parameters, we achieved a precision of 75%, and a recall of up to 60%. Both methods discovered about 60% of all yeast interactions described in the DIP-referenced articles. This task was slightly different from the above, because the DIP nodes do not reference the exact textual evidence, but the complete text only.

Next we learned FSA structures from our corpus using genetic optimization. Setting the fitness function, which

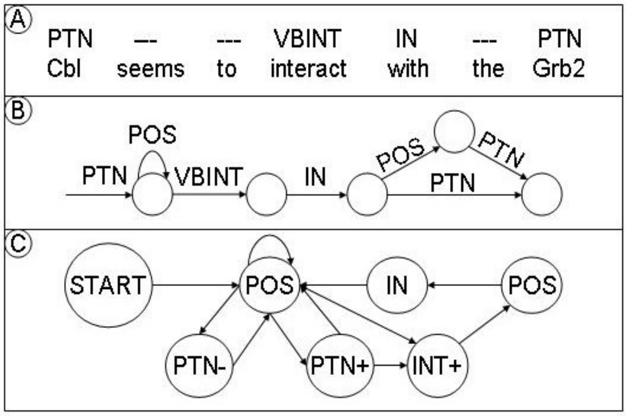
A
PTN --- --- VBINT IN --- PTN
Cbl seems to interact with the Grb2
B
POS
PTN
VBINT
IN
POS
PTN
PTN
C
START
POS
IN
POS
PTN-
PTN+
INT+

Figure 1: A pattern as a sequence of anchor words aligned to a text (A) and as a regular expression modeled by a Mealy-FSA (B). The HMM on the bottom is a probabilistic Moore-FSA (C). We distinguished between related entities (+) and not related entities (-) in annotated corpora and represented both occurrences for each type of entity by a different state. PN - protein name; POS - any POS tag; VBINT - verb referring to an interaction; IN - preposition.

directs the evolution process, to either precision or recall, directly effects structures of emerging patterns. More than 65% of DIP interactions were found using only four recall-optimized Mealy-FSAs. When optimizing patterns toward precision, crossvalidation revealed an overfitting to training corpora. After reducing model complexity, *e.g.* limiting the number of transitions or allowing only forward-connected models, performance improved.

A problem of time complexity arose from the Viterbi-algorihm, that calculates the state sequence for an HMM and a given sentence. This made evolution of HMM structures much slower than for Mealy structures. Restriction to models with only a few states helped to save learning time.

4 Discussion

Our results indicate that automatically learned patterns show good performance for the extraction of PPI from free text and even outperform hand-crafted patterns [Plake *et al.*, 2005; Blaschke and Valencia, 2002]. Blaschke *et al.* reported a recall of 25% on a set of 851 interactions deduced from the DIP. Huang *et al.* aligned patterns for the extraction of PPI and reported a precision of about 80% on a set of 1200 sentences. Building patterns for an IE-system by hand is a time-consuming and thus expensive step. Experts are needed for analyzing texts to find similar passages or phrases, that describe the desired information. In this work we propose automatic pattern learning, which can be tuned toward certain qualtity measures. Our approach relies on annotated, representative text corpora, which are unfortunately rarely available for most extraction tasks. Here, manual annotation again becomes a bottleneck for automatic IE-systems. For IE-tasks in the domain of life science, some corpora, *e.g.* GENIA [Ohta *et al.*, 2002], BioCreAtIvE, BioText [Rosario and Hearst, 2004], and IEPA [Berleant *et al.*, 2003], with annotated entities such as genes, proteins, drugs, and diseases, are available. The IEPA corpus even comes with annotated PPI, BioText with disease-treatment relations.

We currently study the impact of different scoring schemes and gap penalties, as well as word class tags on extraction performance. Incorporating information from chunkers and shallow parsers can help to capture the relevant parts of a sentence. For learning HMM structures, we also consider techniques other than evolution, *e.g.* merging states starting with a fully specified model [Seymore *et al.*, 1999]. For a broader evaluation of our methods, we will also perform the extraction of relations from other corpora (IEPA, BioText).

Acknowledgments

This work is supported by the German Federal Ministry of Education and Research (BMBF) under grant contract 0312705B. The Knowledge Management in Bioinformatics Group is a member of the Berlin Center for Genome Based Bioinformatics (BCB).

References

[Berleant *et al.*, 2003] Daniel Berleant, Jing Ding, and Andy W. Fulmer. Corpus Properties of Protein Interaction Descriptions in MEDLINE. *http://class.ee.iastate.edu/berleant/home/me/cv/papers/corpuspropertiesstart.htm*, 2003.

[BioCreAtIvE, 2004] Christian Blaschke, Alfonso Valencia, Lynette Hirschman, and Alexander Yeh. BioCreAtIvE Challenge Cup. *BMC Bioinformatics*, 2004.

[Blaschke and Valencia, 2002] Christian Blaschke and Alfonso Valencia. The Frame-Based Module of the SUISEKI Information Extraction System. *IEEE Intelligent Systems*, 17(2):14–20, 2002.

[Daraselia *et al.*, 2004] Nikolai Daraselia, Anton Yuryev, Sergej Egorov, Svetlana Novichkova, Alexander Nikitin, and Ilya Mazo. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–611, 2004.

[Friedman *et al.*, 2001] Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. GENIES: A natural-language processing system for the extraction of molecular pathways from biomedical texts. *Bioinformatics*, 17:74–82, 2001.

[Huang *et al.*, 2004] Minlie Huang, Xiaoyan Zhu, Donald G. Payan, Kunbin Qu, and Ming Li. Discovering patterns to extract protein-protein interactions from full biomedical texts. *Bioinformatics*, 2004.

[Ohta *et al.*, 2002] Tomoko Ohta, Yuka Tateisi, Hideki Mima, and Jun'ichi Tsujii. GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. *Proc. HLT*, 73–77, 2002. *http://www-tsujii.is.s.u-tokyo.ac.jp/∼genia/*

[Plake *et al.*, 2005] Conrad Plake, Jörg Hakenberg, and Ulf Leser. Optimizing Syntax Patterns for Discovering Protein-Protein Interactions. *ACM SAC, Bioinformatics track*, Santa Fe, USA, 2005, To appear.

[Rosario and Hearst, 2004] Barbara Rosario and Marti A. Hearst. Classifying Semantic Relations in Bioscience Texts. *ACL*, Barcelona, Spain, 2004. *http://biotext.berkeley.edu/.*

[Seymore *et al.*, 1999] Kristie Seymore, Andrew McCallum, and Ronald Rosenfeld. Learning Hidden Markov Model Structure for Information Extraction. *AAAI MLIE Workshop*, 1999.

[Viterbi, 1967] Andrew J. Viterbi. Error bounds for convolutional codes and an asymtotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–267, 1967.

[Xenarios *et al.*, 2001] Ioannis Xenarios, Esteban Fernandez, Lukas Salwinski, Xiaoqun J. Duan, Michael J. Thompson, Edward M. Marcotte, and David Eisenberg. DIP: the Database of Interacting Proteins: 2001 update. *Nucleic Acids Res*, 29:239–241, 2001.